CULSPIN – Compute ULam SPiral INdices v-1.0

Manual

de

Autores

Lázaro Guillermo Pérez Montoto
Francisco Javier Prado Prado
Cristian R. Munteanu
Humberto González Díaz

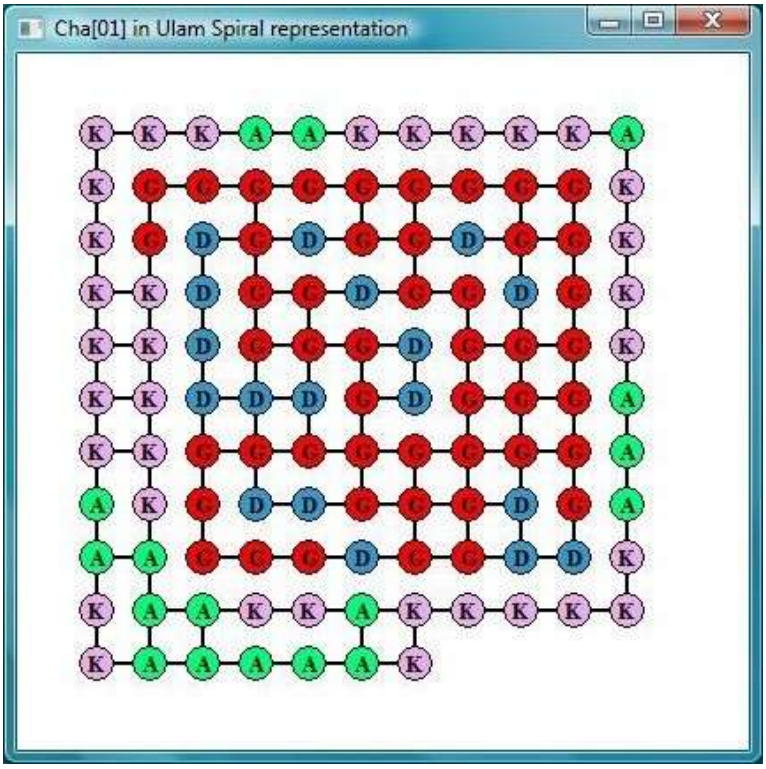
2009

(lgpm2002@yahoo.es)

¿Qué es CULSPIN?

CULSPIN transforma cualquier secuencia de letras en una representación gráfica que usa como plantilla la espiral de Ulam (disposición de los números naturales en forma de espiral) y en la que se conectan aquellos nodos que pertenecen a la misma clase (tienen la misma letra). Por ejemplo:

```
Cha[01]
GDDGGDGGGGGGGGDGGGDGDDGGGDGGGDGDGGDGDDDDGGGGGDGGDDGGGGGGGGGGGGGGGKKKKKAAAKKAKKKKKKAAA
KKKKAKKKKKAAKKKKKKKKKAAKKAAAAAK
```



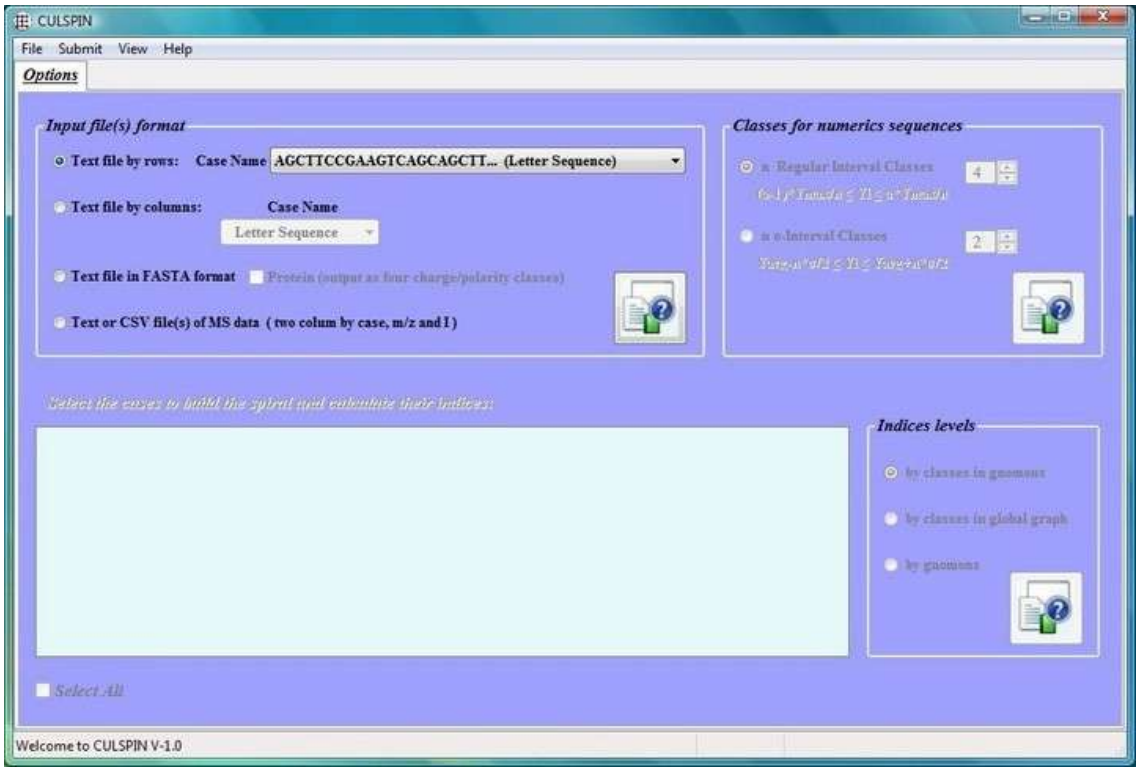
Además, basándose en ese grafo, CULSPIN calcula dos familias de Índices Topológicos (***TIs***). Estos índices pueden ser calculados a varios niveles: para cada una las clases en cada gnomon de Ulam, para cada una de las clases en todo el grafo y para cada gnomon independiente de las clases. Por otra parte, los grafos 2D (***Grafos-U***) generados por la aplicación, además de ser visualizados, pueden ser exportados con el objetivo de poder utilizarlos en otros programas para calcular otras familias de ***TIs***. Todos los índice numéricos se pueden guardar y/o exportados y con ellos se pueden realizar diversos análisis estadísticos o crear modelos QSAR (relación estructura - propiedades). Ejemplos de secuencias son las cadenas de amino ácidos de las proteínas, los ácidos nucleídos y los espectros de masa de proteínas. CULSPIN se puede utilizar para estudiar distintos sistemas, desde los sistemas simples de átomos en moléculas pequeñas anti-cancerígenas, hasta sistemas complejos de redes metabólicas, sociales, computacionales o sistemas biológicos.

¿Qué puede hacer CULSPIN?

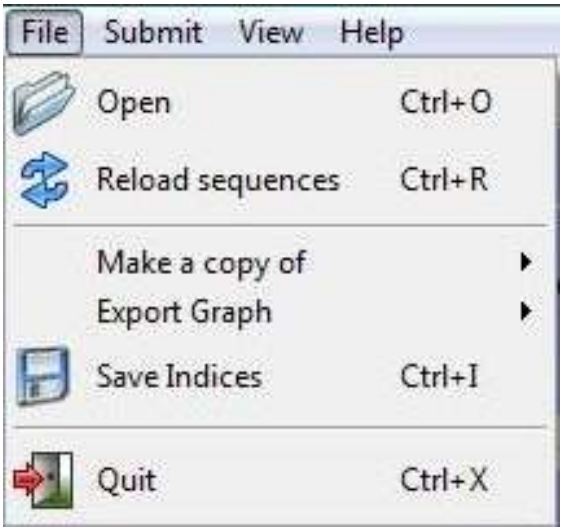
- ***Leer*** secuencias de letras organizadas por filas o columnas a partir de ficheros ***TXT***;
- ***Leer*** secuencias en formato ***FASTA*** almacenadas en ficheros ***TXT***;
- ***Leer*** secuencias o series numéricas, organizadas por filas o columnas a partir de ficheros ***TXT***;
- ***Leer*** datos numéricos correspondientes a señales de Espectros de Masa **(*MS*)** a partir de múltiples ficheros ***TXT*** *o* ***CSV***;
- ***Convertir*** secuencias o series numéricas y datos de ***MS*** en secuencias de letras;
- ***Transformar*** cualquier secuencia de letras en su correspondiente ***Grafo-U*** conectando los nodos que pertenezcan a la misma clase (tienen la misma letra);
- ***Calcular*** dos familias de ***TIs*** usando los ***Grafos-U*** generados y ***Mostrar*** sus valores en una tabla;
- ***Graficar*** y ***Visualizar*** el ***Grafo-U*** de la secuencia que se seleccione*;*
- ***Exportar*** la información de la conectividad de los ***Grafo-U*** en ficheros ***CT*** *o* ***NET****;*
- ***Guardar*** los ***TIs*** calculados en ficheros ***TXT*** *o* ***CSV***.

¿Cómo utilizar CULSPIN?

CULSPIN es una aplicación interactiva creada con Python/wxPython con formato de libreta de notas que presenta una barra de menú principal con las siguientes opciones:



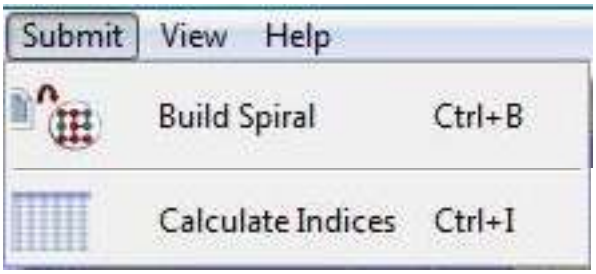
Menú File :



- ***Open file*** **:** permite buscar, seleccionar, abrir el fichero del cual se tomarán los datos de entrada (secuencias de letras, secuencias o series de numéricas, etc.). Una vez cargados los datos, las secuencias de letras se muestran en una lista.
- ***Reload sequences*** **:** permite volver a trabajar con las secuencias cargadas inicialmente (secuencias originales). Esta opción sólo se activa si no se le construyó la espiral a todas las secuencias originales. Un vez terminado el proceso de recarga, todas las secuencias originales vuelven a estar disponibles en la lista.

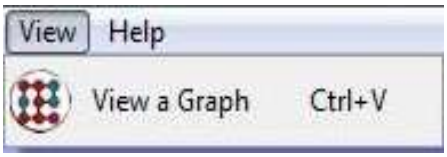
- ***Make a copy of* :** hacer una copia en un fichero ***TXT*** de las secuencias de letras originales o las secuencias de letras estudiadas, pero en el formato con el que se muestran en la lista (***nombre*** <espacio>***secuencia***). Esta opción está disponible o activa sólo cuando las secuencias mostradas al abrir el fichero, han requerido cierta transformación, es decir, cuando los datos estaban organizados por columnas, eran números, estaban en formato ***FASTA***, etc.
- ***Export graph* :** exportar a ficheros independientes de tipo ***CT*** o ***NET*** la conectividad de cada uno de los ***Grafos-U*** construidos, con el objetivo de poder utilizarlos en otros programas para someterlos a otros cálculos.
- ***Save Indices* :** guardar en ficheros ***TXT*** o ***CSV*** los índices calculados por la aplicación para su posterior estudio estadístico
- ***Quit* :** salir de la aplicación.

Menú Submit :



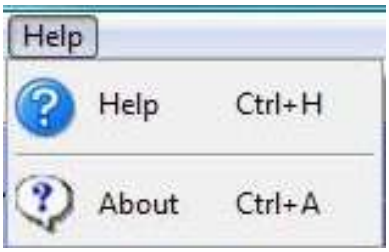
- ***Build Spiral* :** colocar las secuencias seleccionadas en la representación de espiral y construir el ***Grafo-U*** conectando los nodos que pertenecen a la misma clase (los que tienen la misma letra).
- ***Calculate Indices* :** calcular los ***TIs*** de las secuencias seleccionadas a partir de sus respectivos ***Grafos-U***. Una vez terminada esta operación, los resultados se muestran en una nueva página.

Menú View :



- ***View a graph*** **:** graficar y visualizar, en una ventana independiente, el ***Grafo-U*** de una secuencia seleccionada (una secuencia a la vez). Sólo esta activa después de haber construido al menos un ***Grafo-U***.

Menú Help :



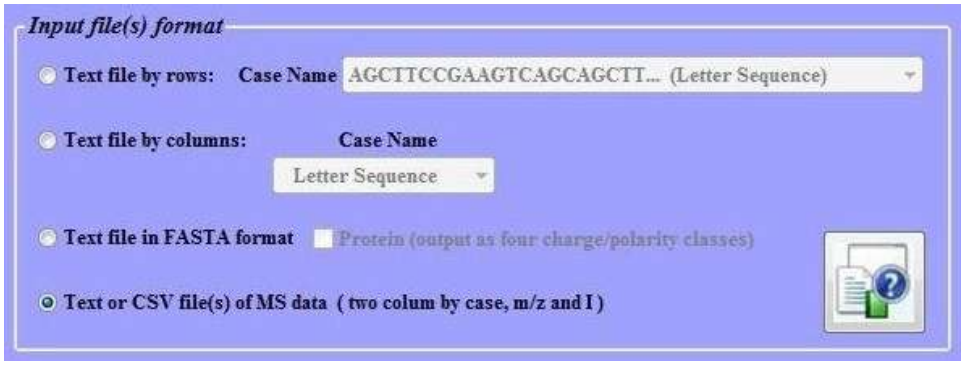
- ***Help*** **:** muestra en una ventana independiente el contenido la ayuda.
- ***About*** **:** muestra la clásica ventana con información acerca de la aplicación.

En un inicio, CULSPIN presenta una sola página con el título ***Options*** en su ventana principal en forma de libreta de notas. Está página tiene como y después de llevar a cabo el cálculo de los ***TIs*** a una o más secuencias, se adiciona un nueva página con el título ***Indices***.

Página Options **:**

En esta página hay cuatro áreas bien definidas cuyas funciones se describen a continuación:

I- Input file(s) format**:** esta caja de controles permite seleccionar, entre los tipos de formatos de ficheros de entrada aceptados por CULSPIN, aquella opción que se corresponda con el formato de nuestros datos.



A continuación mostramos un ejemplo de cada uno de los formatos para su mejor comprensión.

a) Text file by rows*:** en este formato las secuencias están organizadas de forma tal que cada línea del fichero ***TXT corresponde a un caso o secuencia diferente.

Secuencias de letras:

```
Cha[01] GDDGGGGGDGGGDGDDGGGDGGGDGDGGDGDDDDGGGGGDGGDDGGGGGGGGGGGKKKKKAAAKKAKKKKAAK
Cha[02] DDGGDGGGGGGGGDGGGDGDDDDDDGGGGGDGGDDGGGGGGGGGGGGGGGGKKKKKAAAKKAKKKKK
Cha[03] GDGGDGGGGGGGGDGGGDGDDGGGDGGGDGDGGDGDDDDGGGGGDGGDDGGGGGGGGGGKKKKKAAAKKAKKKKKKAAA
```

Secuencias numéricas:

```
Cha[01] -7.86E-05 2.18E-07 9.60E-05 0.000366 0.000810 0.001428 0.002221 0.00318 0.004328
Cha[02] 2.18E-07 9.60E-05 0.000366 0.000810 0.001428 0.002221 0.003187 0.00432 -7.86E-05
Cha[03] 9.60E-05 0.000366 0.000810 0.001428 0.002221 0.003187 0.004328 0.005643
```

***b-) Text file by columns*:** en este formato las secuencias están organizadas de forma tal que cada columna en el fichero texto corresponde a un caso o secuencia diferente.

Secuencias de letras:

```
Cha[01] Cha[02] Cha[03]
D G G
D D D
G D G
G G G
D G D
G D G
G G G
G G G
G G G
G G G
G G G
G G G
```

Secuencias numéricas:

```
Cha[01] Cha[02] Cha[03]
-7.86E-05 2.18E-07 9.60E-05
2.18E-07 9.60E-05 0.00036601
9.60E-05 0.00036601 0.0008102
0.00036601 0.0008102 0.00142856
0.0008102 0.00142856 0.00222112
0.00142856 0.00222112 0.00318787
0.00222112 0.00318787 0.00432881
0.00318787 0.00432881 0.00564393
0.00432881 0.00564393 0.00713324
0.00564393 0.00713324 0.00879674
0.00713324 0.00879674 0.01063443
0.00879674 0.01063443 0.01264631
0.01063443 0.01264631 0.01483238
0.01264631 0.01483238 0.01719263
0.01483238 0.01719263 0.01972708
```

c-) Text file in FASTA format:

```
>gi|221068402|ref|ZP_03544507.1|enzyme [Comamonas testosteroni KF-1]
MSEPVNQWPQTLEERIDRLESLDAIRQLAGKYSLSLDMRDMDAHVNLFAPDIKVGKEKVGRAHFMAWQDS
TLRDQFTGTSHHLGQHIIEFVDRDHATGVVYSKNEHECGAEWVIMQMLYWDDYERIDGQWYFRRRLPCYW
YATDLNKPPIGDMKMRWPGREPYHGAFHELFPSWKEFWAQRPGKDQLPQVAAPAPLEQFLRTMRRGTPAP
RMRVR

>gi|220713425|gb|EED68793.1| enzyme [Comamonas testosteroni KF-1]
MSEPVNQWPQTLEERIDRLESLDAIRQLAGKYSLSLDMRDMDAHVNLFAPDIKVGKEKVGRAHFMAWQDS
TLRDQFTGTSHHLGQHIIEFVDRDHATGVVYSKNEHECGAEWVIMQMLYWDDYERIDGQWYFRRRLPCYW
YATDLNKPPIGDMKMRWPGREPYHGAFHELFPSWKEFWAQRPGKDQLPQVAAPAPLEQFLRTMRRGTPAP
RMRVR

>gi|77360245|ref|YP_339820.1| enzyme [Pseudoalteromonas haloplanktis TAC125]
MQYLVISDIYGKTPCLQQLAKHFNAENQIVDPYNGVHQALENEEEYYKLFIKHCGHDEYAAKLEEYFNKL
SKPTICIAFSAGASAAWRAQASTTTTHLKKVIAFYPTQIRNYLNIDAIHPCEFIFPGFEPHFNVDELITN
LSAKNNVRCLKTLYLHGFMNQQSQNFSEYGYQYFYKVIKTANSEAH
```

Note: *en el caso de las proteínas, si se selecciona la opción* ***Protein*** *cada aminoácido presente en la secuencia se codifica con una letra o clase diferente. Para ello se tiene en cuenta el grupo al que pertenezca el aminoácido según la polaridad y las propiedades ácido-base de sus cadenas laterales:* ***no polar y neutro****;* ***polar y neutro; ácido y polar****; y* ***básico y polar****.*

d-) Text or CSV files of MS data: En esta opción cada caso se encuentra almacenado en un fichero independiente. En ellos los datos de las señales del espectro están organizados en dos columnas: ***masa/carga*** (***m/z***) e ***Intensidad*** con encabezado o no. Los ficheros pueden ser de tipo ***TXT*** o ***CSV***.

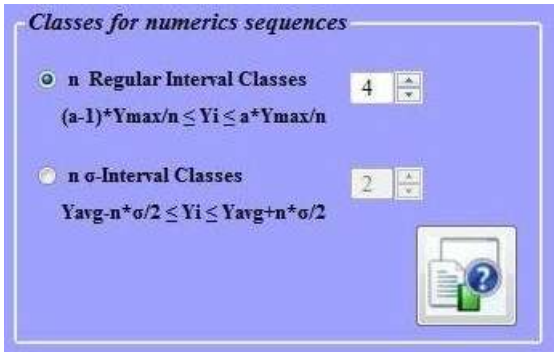
Ficheros TXT: (las columnas están separadas por tabulación)

```
2.5660      0.6601
3.6601      8.9102
8.1024      42.0856
14.2856     22.2112
22.2112     3.8787
31.8787     4.3288
43.2881     56.4393
56.4393     71.3324
71.3324     87.9674
87.9674     90.0000
106.3443    12.1631
126.4631    8.3238
148.3238    100.9263
```

Ficheros CSV: (los elementos están separados por comas)

```
m/Z,Intensity
2.5660,0.6601
3.6601,8.9102
8.1024,42.0856
14.2856,22.2112
22.2112,3.8787
31.8787,4.3288
43.2881,56.4393
56.4393,71.3324
71.3324,87.9674
87.9674,90.0000
106.3443,12.1631
126.4631,8.3238
148.3238,100.9263
```

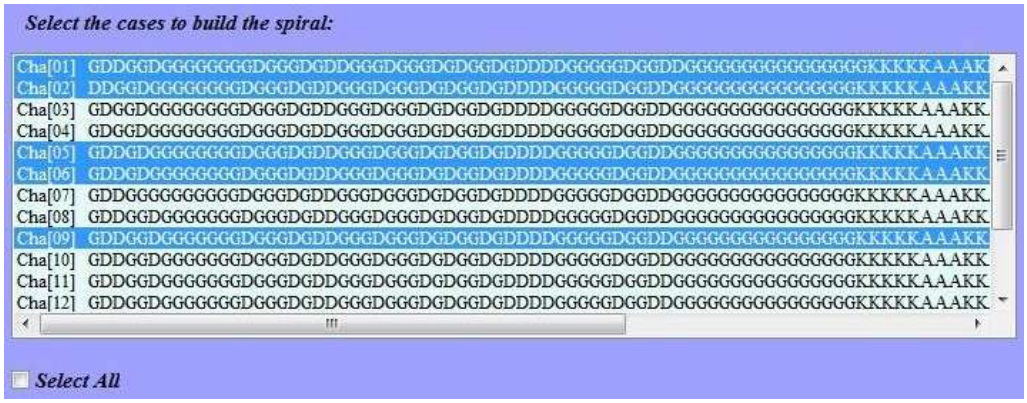
II- Classes for numerics sequences: esta caja de controles sólo está active si el formato de entrada seleccionado es de tipo numérico. En ella se ofrecen dos heurísticas diferentes para transformar una secuencia o serie numérica en una secuencia de letras.



- ***n Regular Interval Classes:*** en esta opción los datos numéricos tomados del fichero de entrada se dividen en **n** intervalos o clases ($2 \leq \mathbf{n} \leq 10$) y se les asigna una letra diferente. Entonces, cada elemento de la secuencia o serie numérica se codifica con la letra de la clase a la que pertenece.
- ***n σ-Interval Classes:*** en esta opción los datos numéricos tomados del fichero de entrada se dividen en $2\mathbf{n}+2$ intervalos ($2 \leq \mathbf{n} \leq 4$) cuyas dimensiones dependen de la desviación estándar de los datos. A cada intervalo o clase se le asigna una letra y se codifica cada elemento de la secuencia o serie numérica con la letra de la clase a la que pertenezca.

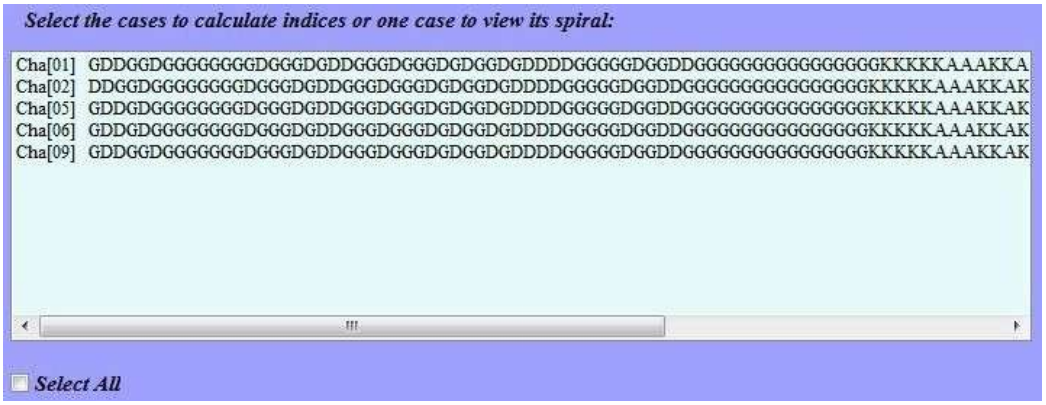
Note: *En el caso de los datos de MS, la presente versión de CULSPIN, los transforma previamente en una serie numérica en la que cada elemento es el producto de la m/z por la intensidad de cada señal del espectro. Luego esta serie numérica es transformada en una secuencia de letras utilizando la heurística seleccionada por el usuario.*

III- A list box for view/select sequences: esta caja de lista tiene la función de mostrar y permitir la selección de secuencias o casos. En un inicio la lista está vacía y después de leer los datos a partir del fichero de entrada, la lista muestra las secuencias leídas directamente del fichero u obtenidas mediante alguna codificación o transformación de las explicadas anteriormente. Una vez que las secuencias de letras son mostradas en esta caja de lista, aparece una invitación a seleccionar las secuencias o casos a los que se les desea construir su ***Grafo-U***.



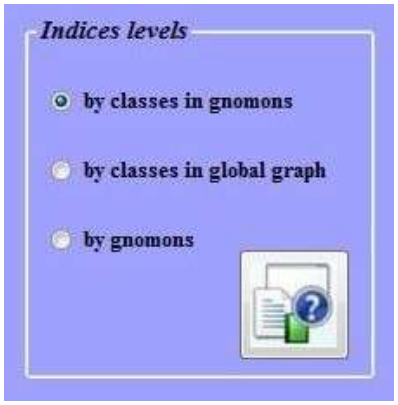
Se puede seleccionar un bloque continuo de secuencias o casos manteniendo presionada la tecla ***Mayúsculas*** al seleccionar el primero y el último caso que conforman el bloque; seleccionar casos alternos regularmente o no presionando la

tecla ***Ctrl*** mientras se seleccionan los casos deseados; o seleccionar todos los casos marcando la opción ***Select All***. Después de construir los ***Grafos-U*** de las secuencias seleccionadas, la caja de lista mostrará sólo los casos con los que se trabajó. En este momento se invita entonces a seleccionar los casos a los que se les desea calcular los ***TIs*** o seleccionar un único caso para ver su grafo en una ventana independiente.



El resto de las secuencias no estudiadas se pueden recuperar sin necesidad de leer nuevamente el fichero de entrada, mediante la opción ***Reload sequences*** presente en el menú ***File***. En tal caso se comienza desde cero, es decir, se perderán los grafos y los ***TIs*** calculados si no se han guardado en ficheros.

IV- Indices levels: esta caja de controles sólo se encuentra activa si se ha construido al menos una espiral y permite seleccionar a qué nivel queremos calcular las dos familias de ***TIs*** implementadas en esta versión de CULSPIN.

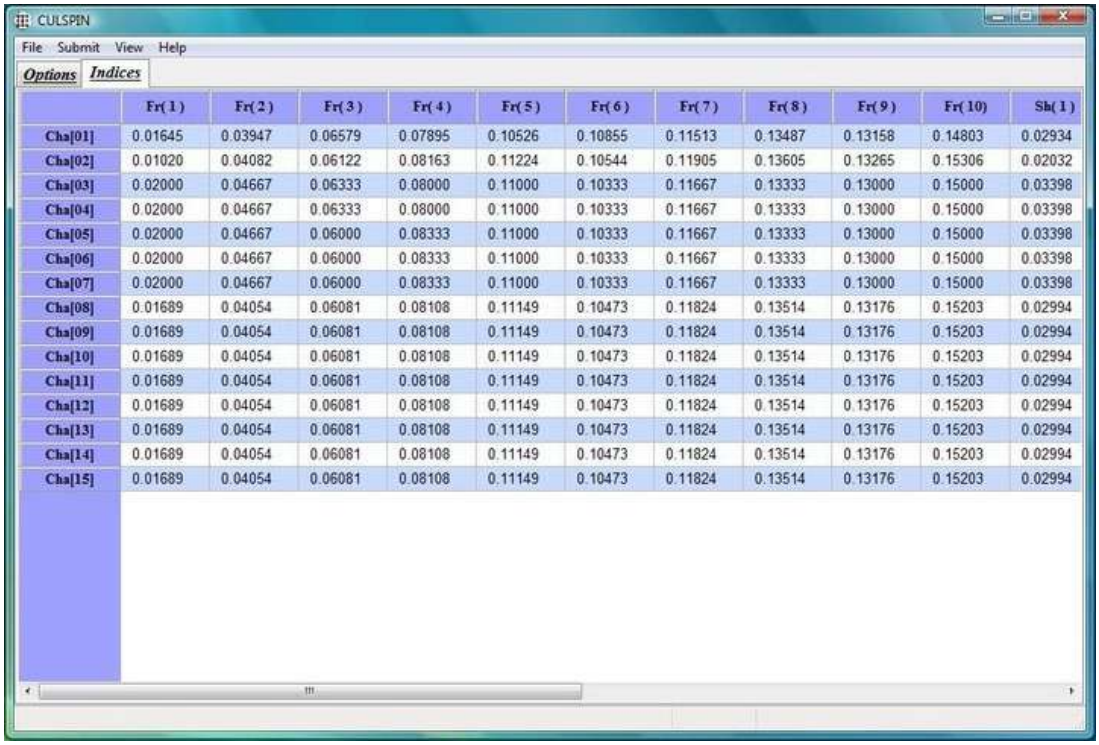


- ***by classes in gnomons:*** si se selecciona esta opción, las dos familias de ***TIs*** se calculan para cada una de las clases en cada uno de los gnomones. En el caso en que una clase no se encuentre en un determinado gnomon, su ***Frecuencia*** y su ***Entropía de Shannon*** en ese gnomon son cero. Esta opción es más útil cuando las secuencias no tienen muchas clases y no son muy grandes, en caso contrario, se obtendría un número demasiado elevado de índices y por tanto su procesamiento estadístico posterior muy engorroso.

- ***by classes in global graph:*** en esta opción los ***TIs*** se calculan para cada una de las clases pero en todo el grafo. En otras palabras, los ***TIs*** de una clase dada en todo el grafo, son el resultado de la sumatoria sus valores en todos los gnómones. Esta opción reduce el número de ***TIs*** en el caso de secuencias muy grandes por lo que resulta una buena opción en tales casos.
- ***by gnomons:*** si se selecciona esta opción, los ***TIs*** se calculan a nivel de gnomones independientemente de las clases. En otras palabras, los índices para un gnomon determinado son el resultado de la sumatoria de los ***TIs*** de todas las clases en ese gnomon. Esta opción puede ser muy útil si se trabaja con secuencias de tamaño moderado y con un gran número de clases.

Página Indices :

Esta página se adiciona a la libreta y se muestra al usuario inmediatamente después de que se calculen los ***TIs*** a las secuencias seleccionadas. El formato de la página es el de una tabla tipo hoja de cálculo, en la que en el encabezado de las columnas se muestran los nombres de los índices y el de las filas el de las secuencias o casos.



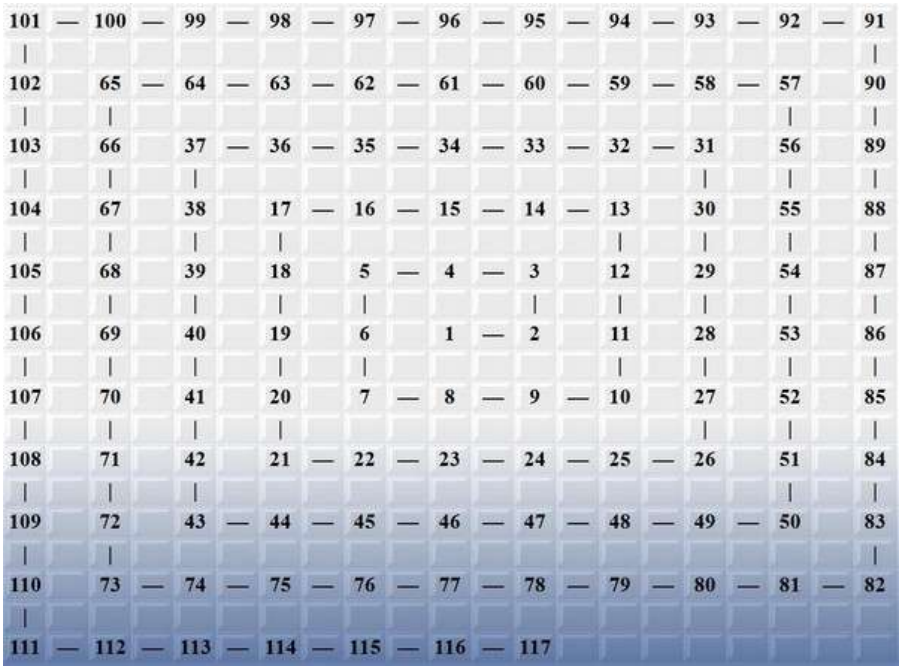
	Fr(1)	Fr(2)	Fr(3)	Fr(4)	Fr(5)	Fr(6)	Fr(7)	Fr(8)	Fr(9)	Fr(10)	Sh(1)
Cha[01]	0.01645	0.03947	0.06579	0.07895	0.10526	0.10855	0.11513	0.13487	0.13158	0.14803	0.02934
Cha[02]	0.01020	0.04082	0.06122	0.08163	0.11224	0.10544	0.11905	0.13605	0.13265	0.15306	0.02032
Cha[03]	0.02000	0.04667	0.06333	0.08000	0.11000	0.10333	0.11667	0.13333	0.13000	0.15000	0.03398
Cha[04]	0.02000	0.04667	0.06333	0.08000	0.11000	0.10333	0.11667	0.13333	0.13000	0.15000	0.03398
Cha[05]	0.02000	0.04667	0.06000	0.08333	0.11000	0.10333	0.11667	0.13333	0.13000	0.15000	0.03398
Cha[06]	0.02000	0.04667	0.06000	0.08333	0.11000	0.10333	0.11667	0.13333	0.13000	0.15000	0.03398
Cha[07]	0.02000	0.04667	0.06000	0.08333	0.11000	0.10333	0.11667	0.13333	0.13000	0.15000	0.03398
Cha[08]	0.01689	0.04054	0.06081	0.08108	0.11149	0.10473	0.11824	0.13514	0.13176	0.15203	0.02994
Cha[09]	0.01689	0.04054	0.06081	0.08108	0.11149	0.10473	0.11824	0.13514	0.13176	0.15203	0.02994
Cha[10]	0.01689	0.04054	0.06081	0.08108	0.11149	0.10473	0.11824	0.13514	0.13176	0.15203	0.02994
Cha[11]	0.01689	0.04054	0.06081	0.08108	0.11149	0.10473	0.11824	0.13514	0.13176	0.15203	0.02994
Cha[12]	0.01689	0.04054	0.06081	0.08108	0.11149	0.10473	0.11824	0.13514	0.13176	0.15203	0.02994
Cha[13]	0.01689	0.04054	0.06081	0.08108	0.11149	0.10473	0.11824	0.13514	0.13176	0.15203	0.02994
Cha[14]	0.01689	0.04054	0.06081	0.08108	0.11149	0.10473	0.11824	0.13514	0.13176	0.15203	0.02994
Cha[15]	0.01689	0.04054	0.06081	0.08108	0.11149	0.10473	0.11824	0.13514	0.13176	0.15203	0.02994

En esta tabla se puede seleccionar una celda, un rango, una columna, un fila o todas las celdas y copiar el contenido de la selección en el clipboard mediante la combinación **Ctrl+C** para luego pegarlo en donde se desee. Esta posibilidad es muy útil si se desea exportar de modo rápido, sencillo y fácil, los valores de los ***TIs*** calculados en aplicaciones externas teles como Excel.

Espiral de Ulam

En 1963 el matemático *Stanisław M. Ulam* descubrió ciertos aspectos interesantes relacionados con la disposición que adoptan los números primos al colocar los números naturales en forma de una espiral. Luego esta disposición tomó mucho auge en la generación y visualización de imágenes.

Para construir la espiral se colocan los números en una rejilla de cuadrículas comenzando por 1 en el centro y luego los demás formando una espiral cuadrada según la siguiente figura:



En matemáticas, esta representación es un método simple de graficar números con el que se revelen aspectos ocultos y muy interesantes de las series y secuencias numéricas. En el estudio de las moléculas, esta representación en espiral ha sido asociada en muchos trabajos encaminados a representar secuencias de nucleótidos de ADN divididos en cuatro clases (A,T,G y C).

¿Qué es un gnomon?

La espiral de Ulam puede dividirse en diferentes regiones o intervalos nombrados gnomones o disposiciones angulares según se puede observar en la siguiente figura:

101	100	99	98	97	96	95	94	93	92	91
102	65	64	63	62	61	60	59	58	57	90
103	66	37	36	35	34	33	32	31	56	89
104	67	38	17	16	15	14	13	30	55	88
105	68	39	18	5	4	3	12	29	54	87
106	69	40	19	6	1	2	11	28	53	86
107	70	41	20	7	8	9	10	27	52	85
108	71	42	21	22	23	24	25	26	51	84
109	72	43	44	45	46	47	48	49	50	83
110	73	74	75	76	77	78	79	80	81	82
111	112	113	114	115	116	117				

. Para definir un gnomon es necesario recordar los números oblongos que son aquellos que se pueden representar mediante el producto n(n+1) con n natural, es decir: 2, 6, 12, 20, 30, 42, 56, 72, 90,.... Estos números dividen a los números naturales en intervalos crecientes en longitud (2n). Resulta fácil de ver que un par de números oblongos consecutivos definen un gnomon y que estas disposiciones angulares se van encajando dando lugar a rectángulos de magnitud creciente. Además queda claro que cada elemento de la espiral pertenece a un único gnomon, es por ello que se puede definir la coordenada ***U*** de un elemento en la espiral de Ulam como el número del gnomon al que pertenece.

Cuando se representa una secuencia de letras en su ***Grafo-U***, cada nodo es un elemento de la secuencia cuya letra representa la clase a la que pertenece dicho elemento y en cada gnomon existirán una o más clases diferentes.

K	K	K	A	A	K	K	K	K	K	A
K	G	G	G	G	G	G	G	G	G	K
K	G	D	G	D	G	G	D	G	G	K
K	K	D	G	G	D	G	G	D	G	K
K	K	K	A	A	K	K	K	K	K	A
K	G	G	G	G	G	G	G	G	G	K
K	G	D	G	D	G	G	D	G	G	K
K	K	D	G	G	D	G	G	D	G	K
A	A	G	G	G	D	G	G	D	D	K
K	A	A	K	K	A	K	K	K	K	K
K	A	A	A	A	A	K				

Índices, definición y cálculo

Como se ha comentado desde un inicio, en los ***Grafos-U*** construidos con ayuda de CULSPIN, cada nodo pertenece a una clase determinada y ellos no sólo están conectados siguiendo la secuencia de letras, sino que además aquellos nodos que pertenecen a la misma clase (tienen igual letra) se conectan entre sí. De modo que, en nuestros ***Grafos-U*** cada nodo estará conectado con uno o más nodos. Por definición, se conoce como ***grados*** de un nodo al número de nodos con los que está conectado el nodo en cuestión y por ***grados totales*** de un grafo a la suma de los grados de todos los nodos que conforman el grafo, entonces podemos definir como grados de un gnomon a la suma de los grados de los nodos que pertenecen a dicho gnomon.

Teniendo en cuenta todo lo anterior, los índices calculados por CULSPIN se definen y calculan del siguiente modo:

Indices levels	Frequency	Shannon Entropy
by classes in gnomon	$f(c,g)=\sum_{\substack{n\in c\\ n\in g}}\frac{\deg(n(c,g))}{\sum_{i\in G_c}\deg(i)}$ c: class; g: gnomon; $n_{c,g}$: node with class c in gnomon g	$Sh(c,g)=-f(c,g)\log(f(c,g))$
by classes in global graph	$f(c)=\sum_{\substack{n\in c\\ n\in G_U}}\frac{\deg(n(c))}{\sum_{i\in G_c}\deg(i)}$ c: class; n_c: node with class c in G_U	$Sh(c)=-f(c)\log(f(c))$
by gnomons	$f(g)=\sum_{n\in g}\frac{\deg(n(g))}{\sum_{i\in G_c}\deg(i)}$ g: gnomon; n_g: node in gnomon g	$Sh(g)=-f(g)\log(f(g))$